下载APP

春节加餐2 | 一次答疑，带你回顾模型评估的所有基础概念

2021-02-15 刘海丰

成为AI产品经理 进入课程 >

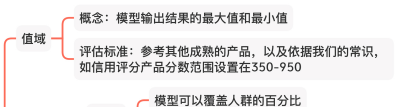
你好，我是海丰。

今天是我们春节假期的第二篇加餐。这节课，我筛选出了一些常见的课后问题，对它们做了统一的整理，同时也对一些优质回答做了补充讲解，希望更多同学可以进行参考。

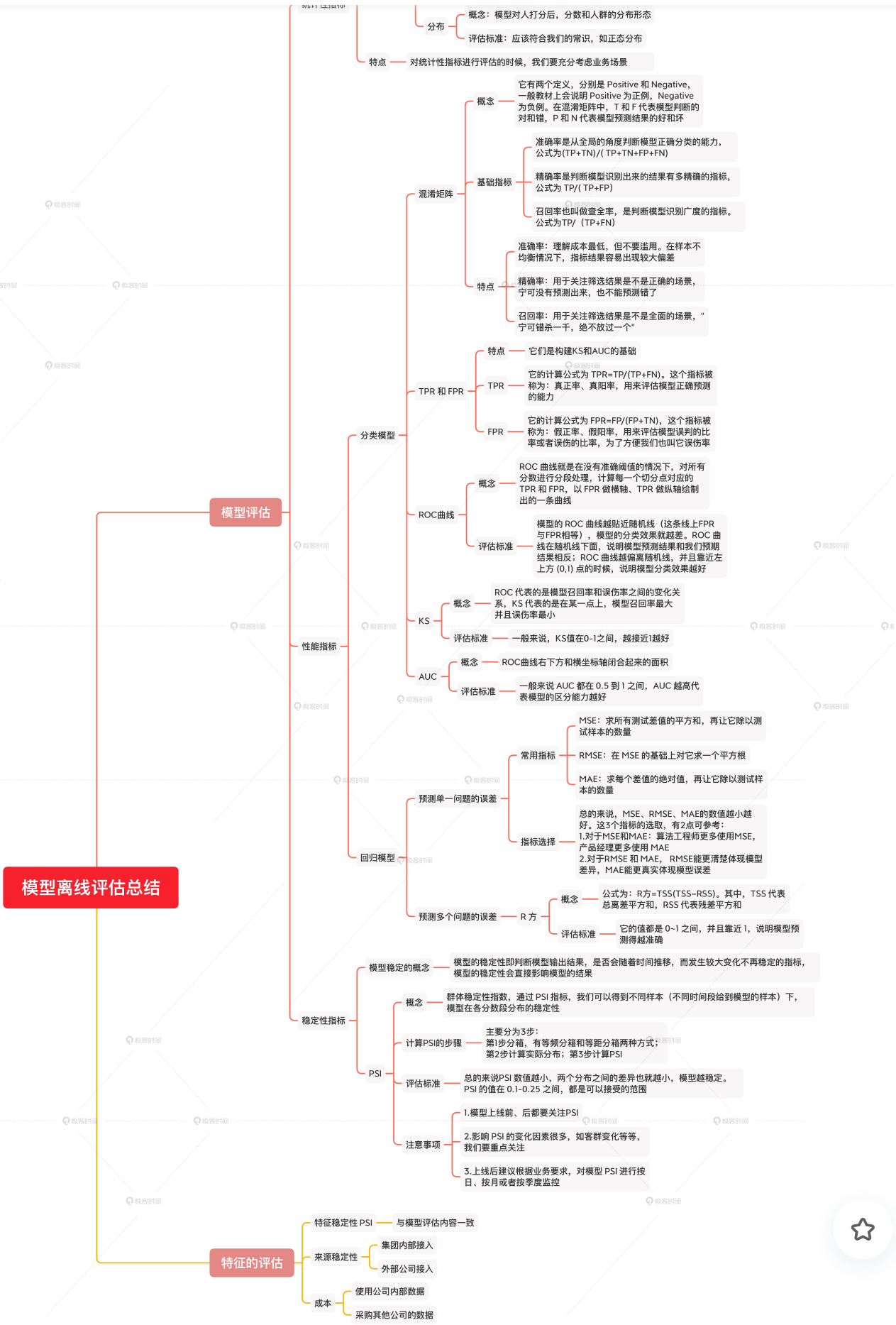
常见问题

问题 1：老师，我总是容易弄混 KS、PSI 等很多指标。

我看到有很多同学提出类似的问题，这些都是 AI 产品经理需要掌握的基础概念，而且在接下来的项目实践篇都会涉及。所以这里，我会把一些基础概念进行汇总，希望能再帮你巩固一下。



下载APP


模型离线评估总结
模型评估
统计性指标
分布
概念：模型对人打分后，分数和人群的分布形态
评估标准：应该符合我们的常识，如正态分布
特点
对统计性指标进行评估的时候，我们要充分考虑业务场景
性能指标
分类模型
混淆矩阵
概念
它有两个定义，分别是 Positive 和 Negative，一般教材上会说明 Positive 为正例，Negative 为负例。在混淆矩阵中，T 和 F 代表模型判断的对和错，P 和 N 代表模型预测结果的好和坏
基础指标
准确率是从全局的角度判断模型正确分类的能力，公式为(TP+TN)/(TP+TN+FP+FN)
精确率是判断模型识别出来的结果有多精确的指标，公式为 TP/(TP+FP)
召回率也叫做查全率，是判断模型识别广度的指标。公式为TP/（TP+FN)
特点
准确率：理解成本最低，但不要滥用。在样本不均衡情况下，指标结果容易出现较大偏差
精确率：用于关注筛选结果是不是正确的场景，宁可没有预测出来，也不能预测错了
召回率：用于关注筛选结果是不是全面的场景，"宁可错杀一千，绝不放过一个"
TPR 和 FPR
特点
它们是构建KS和AUC的基础
TPR
它的计算公式为 TPR=TP/(TP+FN)。这个指标被称为：真正率、真阳率，用来评估模型正确预测的能力
FPR
它的计算公式为 FPR=FP/(FP+TN)，这个指标被称为：假正率、假阳率，用来评估模型误判的比率或者误伤的比率，为了方便我们也叫它误伤率
ROC曲线
概念
ROC 曲线就是在没有准确阈值的情况下，对所有分数进行分段处理，计算每一个切分点对应的 TPR 和 FPR，以 FPR 做横轴、TPR 做纵轴绘制出的一条曲线
评估标准
模型的 ROC 曲线越贴近随机线（这条线上FPR与FPR相等），模型的分类效果就越差。ROC 曲线在随机线下面，说明模型预测结果和我们预期结果相反；ROC 曲线越偏离随机线，并且靠近左上方 (0,1) 点的时候，说明模型分类效果越好
KS
概念
ROC 代表的是模型召回率和误伤率之间的变化关系，KS 代表的是在某一点上，模型召回率最大并且误伤率最小
评估标准
一般来说，KS值在0-1之间，越接近1越好
AUC
概念
ROC曲线右下方和横坐标轴闭合起来的面积
评估标准
一般来说 AUC 都在 0.5 到 1 之间，AUC 越高代表模型的区分能力越好
回归模型
预测单一问题的误差
常用指标
MSE：求所有测试差值的平方和，再让它除以测试样本的数量
RMSE：在 MSE 的基础上对它求一个平方根
MAE：求每个差值的绝对值，再让它除以测试样本的数量
指标选择
总的来说，MSE、RMSE、MAE的数值越小越好。这3个指标的选取，有2点可参考：
1.对于MSE和MAE：算法工程师更多使用MSE，产品经理更多使用 MAE
2.对于RMSE 和 MAE， RMSE能更清楚体现模型差异，MAE能更真实体现模型误差
预测多个问题的误差
R 方
概念
公式为：R方=TSS(TSS-RSS)。其中，TSS 代表总离差平方和，RSS 代表残差平方和
评估标准
它的值都是 0~1 之间，并且靠近 1，说明模型预测得越准确
稳定性指标
模型稳定的概念
模型的稳定性即判断模型输出结果，是否会随着时间推移，而发生较大变化不再稳定的指标，模型的稳定性会直接影响模型的结果
PSI
概念
群体稳定性指数，通过 PSI 指标，我们可以得到不同样本（不同时间段给到模型的样本）下，模型在各分数段分布的稳定性
计算PSI的步骤
主要分为3步：
第1步分箱，有等频分箱和等距分箱两种方式；
第2步计算实际分布；第3步计算PSI
评估标准
总的来说PSI 数值越小，两个分布之间的差异也就越小，模型越稳定。PSI 的值在 0.1-0.25 之间，都是可以接受的范围
注意事项
1.模型上线前、后都要关注PSI
2.影响 PSI 的变化因素很多，如客群变化等等，我们要重点关注
3.上线后建议根据业务要求，对模型 PSI 进行按日、按月或者按季度监控
特征的评估
特征稳定性 PSI
与模型评估内容一致
来源稳定性
集团内部接入
外部公司接入
成本
使用公司内部数据
采购其他公司的数据

极客时间

> 在我刚开始转做 AI 产品经理的时候，遇到过这么一件事儿，我们的用户年龄预测模型训练时候 的 KS 值（模型中用于区分预测正负样本分隔程度的评价指标）很高，但是 OOT 测试的时候 KS 还不足 10。当我拿着结果去找算法同学沟通的时候，他就说了一句，“可能是过拟合了，我再改改”。

也有同学提问说，“不明白为什么 OOT 过高就是模型样本选择的问题？”我这里表达的意思，其实不是 OOT 过高就是模型样本选择有问题，而是说，如果模型在验证集上的 KS 值不错，但到了 OOT 测试的时候 KS 反而不好，就说明，可能是我们当时选择的样本不合理。比如：训练和验证样本抽取的时间过早，很多数据发生了变化，这就导致模型在过去样本上表现好，但在目前样本上不适用。

课后讨论题

问题 2：假设，我们业务方 HR 团队希望你分析每一个员工的离职概率，你觉得用朴素贝叶斯合不合适？为什么？（12 | 朴素贝叶斯：让 AI 告诉你，航班延误险该不该买？）

下载APP

Geek_542a30

离职率是一个非常复杂的模型，朴素贝叶斯算法过于简单了，我认为不太合适

写于 2021/01/08

成为AI产品经理

12 | 朴素贝叶斯：让AI告诉你，航班延误险

识别二维码打开原文
「极客时间」 App

不同，选择的模型不同。因此，我认为对于这个问题，我们可以分为两个阶段来考虑。

第一个阶段：快速响应，从 0 到 1。我们要结合成本来考量，因为一般公司内部部门的诉求样本比较少，并且他们对准确率的接受程度比较大，他们在意的是从无到有。那么，针对启动期没有效果指标可以参考的情况，我们可以先用朴素贝叶斯来完成业务方从无到有的需求，做到快速响应。

第二个阶段：根据实际情况，考虑资源倾斜。这个阶段，我们就要看模型的实际使用情况。不过，一般来说，内部部门的项目就到此为止。但如果你是在众包机构工作，就需要切实投入更多的资源，来训练一个模型解决资源倾斜的问题。

希望同学们可以根据我提供的思路，来继续深入思考一下这个问题。

问题 3：假设，你们公司是做电商业务的，现在领导打算让你做流失用户的召回，你认为评估这个模型的指标都有哪些？为什么？（18 | 核心技能：产品经理评估模型需要关注哪些指标？）

悠悠

一、定义什么算流失用户，确定指标

消费金额 消费频率 登录频次

金额、消费频率应该是最重要的一个指标，因为电商的存在意义就是消费）

二、判断流失用户

是二分类问题，要么是流失用户，要么不是流失用户。可以采用一些分类算法建立模型，来判断是否流失用户。

三、评估模型

要尽可能多找出流失用户进行召回，把部分非流失用户判断成流失用户也是可以的。

要提高召回率，

下载APP

准确率相对降低也没关系。

准确率 = 预测对的 / 所有

召回率 = 预测对的流失用户 / 所有流失用户

——— 写于 2021/01/29

成为AI产品经理
18 | 核心技能：产品经理评估模型需要关注

识别二维码打开原文
「极客时间」 App

这位同学的思路很好！下面，我们再来说一下我的思路。对于产品经理来说，我们首先要明确做一件事情的目的是什么，领导预期达到的目标是怎么样的，然后明确指标的计算口径。

后会流失）。总的来说，分类问题还是使用 KS 作为评估指标比较多。

最后，就是模型的评估。一般来说，我们给到业务评估指标会使用精确率和召回率，这是一对经常被同时使用的指标。这里，我还要提醒你一点，精确率和准确率是两个不一样的指标，它们的概念，我都总结在了刚才的脑图中，我们要记住。除此之外，因为流失用户在整个用户群体中占比不会太大，所以我们一般不使用准确率作为评估指标。

最后，春节假期就快结束了，我也希望这次加餐能帮助你做好准备，迎接新的挑战，我们节后再见！

提建议

更多课程推荐

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

下载APP

精选留言

写留言

> 由作者筛选后的优质留言将会公开显示，欢迎踊跃留言。